Project	CHIL – IP506909 – Computers in the Human Interaction Loop
Title	**Speaker Localization and Tracking - Evaluation Criteria**
Workpackage	WP4 - Speaker Localization and Tracking – Evaluation Criteria
Classification	Draft
Dissemination level	CC: Confidential to the CHIL Consortium
Version	5.0
Date	January 18th, 2005
Number of pages	26
Document ID	CHIL-IRST_SpeakerLocEval –V5.0-2005-01-18-CC
Partners	ITC-irst
Authors	Maurizio Omologo, Alessio Brutti, Piergiorgio Svaizer
Contributors	See above + Luca Cristoforetti, Paolo Coletti
Final editing	Maurizio Omologo
Synopsis	Evaluation Criteria for the Speaker Localization and Tracking Task
Key words	Evaluation, reference transcriptions, labeling, speaker localization, tracking, microphones, time delay estimation (TDE), calibration, speech activity detection (SAD), common sensor set-up, NIST MarkIII array, T-shaped array, video camera

Revision history:

Version	Date	Changes	Editor
1.0	2004-08-12	First Version	M. Omologo
2.0	2004-09-20	Modifications in Sections 2.1, 3.1, 3.2.1, 4.4,5.3 based on a first feedback of the partners involved in activities on speaker localization and tracking	M. Omologo
3.0	2004-10-21	Modifications in Section 5 according to the proposed method for calibration of the microphone network	M. Omologo
4.0	2004-11-30	Minor modifications in Section 4 for what regards the evaluation in frames where a noise source was included in labeling. Appendix A and Appendix B.	M. Omologo
5.0	2005-01-18	Other minor modifications	M. Omologo

Table of contents

1. Introduction

1.1 Objectives of the document

The present document has the objective of describing the criteria adopted in CHIL for the evaluation of speaker localization and tracking technologies. On the basis of the given evaluation criteria, a set of related software tools was developed with the purpose of comparing performance provided by the given technologies, although applied to different experimental contexts. As discussed in [1], in fact, different partners in CHIL will contribute to collect audio-visual corpora under different acoustic conditions, room characteristics, and scenarios. Hence, a coherent method across the laboratories in experimental setting-up, data collection as well as data labeling, is necessary to eventually draw significant conclusions from this research activity.

The next section reports on a brief state of the art of the speaker localization and tracking problem. Section 3 introduces to the experimental contexts in CHIL, Section 4 addresses the evaluation methods here adopted and the related tools. Then, Section 5 illustrates the convenience of a system calibration procedure.

Finally, Appendix A provides some details on the software developed to process XML-based labeling files, while Appendix B reports on an example of application of the given tools to some real lecture data.

It is worth noting that the document is mainly referred to the case of lecture scenario. As a matter of fact, the evaluation software released with the present document has been checked only on lecture data, as at this moment no meeting data have been collected in CHIL. A more detailed discussion on the most convenient criteria to adopt in evaluating speaker localization systems for meeting scenarios will be addressed in a companion document to complete during the second year of CHIL project.

2. Speaker Localization and Tracking

2.1 Problem Definition

Research on Acoustic Source Localization in CHIL refers to locate, identify and track active talkers in enclosures such as offices or meeting rooms. Therefore, we are in presence of large-band, unstationary acoustic emitters acting in closed space of small dimension in relation with the involved wavelengths. Moreover the wave propagation is characterized by reflections on the surfaces and scattering by the objects inside the rooms. The speakers can be modeled as multiple directional acoustic emitters possibly moving in space and overlapping in time. All these aspects make the problem of speaker localization inside rooms a special case in the general topic of passive source localization by means of multiple sensors [2].

A lot of literature exists on the general topic also reporting on methods that could not find direct application in the talker localization scenario (correlation-based and autoregressive methods, eigenvalue-based analysis, MUSIC algorithm), in particular techniques either requiring a priori knowledge on the statistics of the emitters and the background noise, or requiring narrowband signals, or making assumption of far-field and low-reverberation.

On the other hand, for CHIL scenarios the most suitable methods to use seem to be those based on the estimation of time delays, as discussed in the next section.

2.2 TDE Methods

Techniques based on Time Delay Estimation (TDE) and Time Difference Of Arrival (TDOA) at multiple microphones have been shown to be capable of accurate speaker localization even in relatively noisy and reverberant environments. The Phase Transform (PHAT) is a Generalized Cross Correlation (GCC) [3] that has be shown to be a particularly robust TDE technique in presence of reverberation [4, 5].
An improved technique to estimate time delays consists in the analysis of the multichannel spatial correlation matrix, that takes advantage of the redundancy among multiple microphones to reduce the effects of noise and reverberation [6].

For what concerns source coordinate computation, in general some sets of relative delays between microphone pairs are estimated and used to derive the source position that is in the best accordance with them and with the given geometry.

Although one can find many examples in the literature regarding speaker localization and tracking, where performance evaluation criteria are based on the measurement of errors in the time delay axis, due to the diversity of proposed techniques in CHIL the evaluation will directly refer to the errors in the speaker position coordinates, as discussed in the following.

2.3 Acoustic Source Localization Accuracy

The accuracy of a speaker localization system is affected by many factors: the number of exploited microphones, their sensitivity, their spatial and spectral response, their relative geometric position, their distance from the speaker. A crucial aspect that deserves consideration is the possibility to have a direct and unobstructed path between the source (in this case the speaker's mouth) and at least some of the microphones, so that the TDOA of the direct wavefront can be measured at different points in space. The reliability of a time delay estimate depends on the spatial coherence of the acoustic signal reaching the sensors and is affected by the distance between the microphones, the level and spatial coherence of the background noise, and the extent of the room reverberation time.

Another critical aspect that will be addressed in the next sections, regards the dependence of results from reference labelings, which can be strongly influenced by a preliminary calibration step (so far accomplished by using image processing techniques).

2.3.1 Source Localization in Two Dimensions

A single delay estimated between the signals of two microphones determines a surface (hyperboloid) of potential source position in the three-dimensional space. The surface can be reasonably approximated by a cone for distant sources. When multiple delay estimates are derived from multiple microphone pairs, the “best intersection point” (according to a proper definition of a distance measure and a consequent minimization) is assumed as estimated candidate source position. A linear array allows source localization except for a rotation along the array axis. If the height of the source is assumed to be known, the linear array is sufficient for a two-dimensional localization.

2.3.2 Source Localization in Three Dimensions

When a three-dimensional localization is requested, the array geometry should span all the three axes of a cartesian coordinate system. In this case, sub-arrays at different places and with different orientations inside the room have to be used in order to provide an adequate coverage of the possible speaker’s positions. In CHIL, the choice of T-shaped subarrays allows to determine azimuth and elevation angles relative to each subarray. Merging information from different subarrays should lead to a source localization in terms of (x,y,z) coordinates.

However, given a high number of microphones at the same height, the evaluation of z coordinate might be biased by less representative input data. As a consequence, the z-coordinate is here considered as a less relevant feature with respect to x and y coordinates.

3. Acoustic source localization tasks in CHIL

3.1 Introduction

In CHIL, the speaker localization and tracking problem is addressed with the specific purpose of developing technologies having effectiveness in the given lecture and meeting scenarios, that is with speakers in real environments, at no more than 4-5 meters from microphones.

In the literature that regards acoustic source localization and tracking, one can find some arguable evaluation criteria, often referring to simulation experiments (for instance, based on the use of the image method, or based on previously computed real impulse responses), to a single speaker or acoustic source and to very precise localization reference requirements. On the other hand, in CHIL one has to deal with real data and with possible competitive speakers.

For the first seminars collected in CHIL during spring-summer 2004, the lecturer reference coordinates were derived from video recordings. However, the resulting coordinates did not refer to other speakers than the lecturer, this way leading to an unreliable labeling for situations in which the speaker was in the audience. On one hand, we can not penalize localization systems able to detect minor but true acoustic events. On the other hand, in some cases determining in advance the real position of the speaker or of any other acoustic source for accurate labeling purposes will be very difficult or impossible (e.g. the coordinates of a person in the audience, who coughs or produces any other short acoustic event from a position that can not be determined exactly by any of the available video recordings).

As a consequence, here a first issue is to distinguish *Accurate* localization, possible only when very accurate coordinates of the speaker (e.g. the lecturer facing the audience) are available time-frame by time-frame, from *Rough* localization, possible when an acoustic event can not be described with high accuracy (even with a visual inspection of the given lecture recordings), but it can be associated to a specific person in a given area of the room. In the former case, the accurate localization task would deal with the automatic computation of a set of coordinates, with possible fine errors (see the next section) of the order of a few tens of centimeters (actually the error target will depend on the calibration step, as discussed in section 5). On the other hand, in the latter case we can think to a task that deals with the best tolerance of the order of 1 meter (just to give a rough idea).

We felt that distinguishing among the two tasks could also make the activity of labelling CHIL audio corpora easier, at least for what concerns the acoustic events regarding competitive speakers and other odd events that are produced in barely identifiable positions[1]. Finally, note that the latter events could be detected by a Speech Activity Detection (SAD) system and eventually processed by the automatic localization systems; hence, having a reference labelling of their localization is necessary in order to manage fairly the detection and localization of possible false alarms produced by a given localization system.

[1] For a general discussion, here we keep the distinction between *Accurate and Rough Localization*, although in practice the manual labellers will try to provide the most accurate description of any event, even occurring in the audience: only when the speaker can not be identified and located , they will use the label "Unrecognized speaker". The evaluation software was conceived to operate also when this distinction is removed.

3.2 Common Sensor Set-up and Localization Tasks

As reported in the document [1], different experimental contexts are considered across the laboratories of CHIL partners. However, a Minimum Common Sensor Set-Up for the CHIL rooms has been defined as shown in *Figure 1* where, in particular, one can see the presence of three T-shaped microphone array and of the NIST MarkIII array.

The evaluation criteria and related tools were conceived to have comparable applicability, behavior, and performance across the different experimental contexts. Hence, in the following we will refer to that Common Sensor Set-Up, although most of the partners may produce extra data which could eventually be used to improve system potential and related performance.

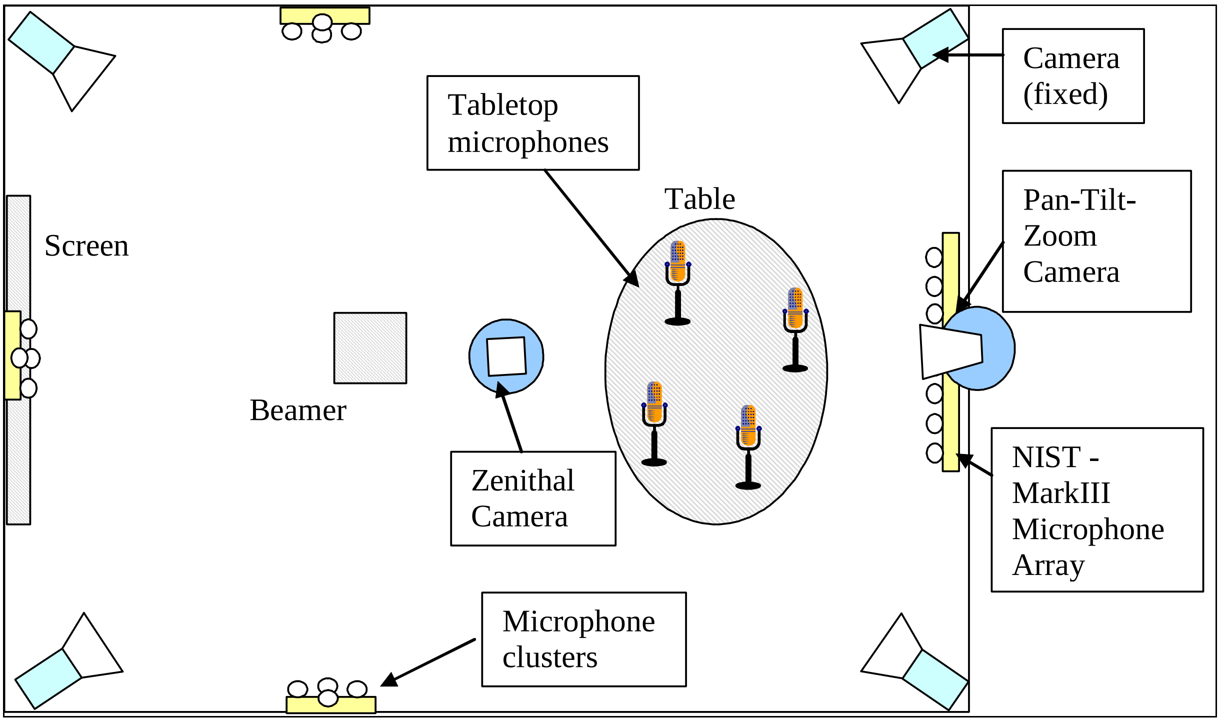

Figure 1: Proposed configuration of a CHIL room with a minimum common sensor set-up: 4 fixed cameras, 1 pan-tilt-zoom camera, 1 zenithal camera, 1 NIST Mark III microphone array, 3 microphone clusters, 4 table-top microphones (close-talking microphones for the speakers are not depicted here)

According to the above mentioned set-up, speaker localization and tracking can be based on processing of the NIST Mark III microphone array signals as well as of the three T-shaped microphone array signals (see the next paragraph). In a meeting scenario context, table-top microphones could be used as well for speaker localization; however, due to the fact that table microphones may be different across the laboratories and that characterizing their positions and use in time would be more difficult (microphones can be moved, or an object can unintentionally be placed as obstacle in front of a microphone during the meeting, or distortions can occur due to vibrations of the table, etc), for the moment <u>we define that the</u>

speaker localization task for meeting scenario will be based on the use of all the far microphones in the minimum common sensor set-up. More specific criteria are subject of a next discussion[2]. Beside the use of the common sensor set-up, another task is foreseen, when an extended sensor set-up is available, based on the use of all the microphones (except table and close-talking ones) available in the experimental room. Results obtained in the latter way might be useful to understand the potential of increasing the number of microphones in the given experimental contexts.

3.2.1 Microphone arrays

The CHIL room includes a NIST-Mark III microphone array. This linear array consists in 64 microphones and has a length of 126 centimeters (based on a distance of 2 cm between adjacent microphones). In the case of a seminar it should be positioned opposite to the presentation area approximately at 4-5 meters from the lecturer. Hence, during a seminar it is expected that the array will pick-up a moderately reverberant voice of the speaker as well as rather varying (in dynamics) voices from the audience (it will depend on the position and head orientation of the latter speakers).

Beside the NIST-Mark III array, other microphone clusters are going to be used in CHIL. The so-called T-shaped microphones are conceived to help in localizing the position of the lecturer, of the other speakers, and of other possible acoustic sources. It is worth noting that microphones of a T-shaped array may acquire a voice from the audience with a very high dynamics, although the NIST-Mark III array is picking up a rather attenuated replica of the same voice. As the objective of a localization system is to estimate the position of any acoustic event, this fact implies that in some situations labeling a NIST-array signal may be not sufficient for the evaluation of the speaker localization system (on the other hand, manual labeling of one microphone of each array would not be feasible).

4. Evaluation criteria and tool development

4.1 Introduction

To summarize, in a lecture we have to locate either the lecturer or a person of the audience, while in a meeting we have to locate any speaker who is talking.

There will be two types of localization tasks, in terms of system accuracy: one, "*Accurate*", corresponding to the situations for which an extremely reliable coordinate reference is available; the other, "*Rough*" used to face with situations for which an accurate localization error evaluation is not possible but it is likely that a system detects and locates an event in the room.

[2] A preliminary distinction in three possible tasks (see previous drafts of this document) was considered too detailed for the purpose of this document, given the fact that no meeting data has been so far collected.

In all of the cases, the speaker localization algorithms can be applied either 1) (just in the case of a meeting) to far microphones and table microphones, or 2) to far microphones of the common sensor set-up, or 3) to all the available far microphones.

4.2 Type of errors

As highlighted above, the localization algorithm will yield a set of coordinates related to the speaker position estimate. Performance will be evaluated by means of the Euclidean distance applied to the coordinates provided by the localization system ($\mathbf{P_l}$) and the corresponding reference coordinates checked by the manual transcriber ($\mathbf{P_r}$).

Localization errors will be classified in two classes [6],[7]:

- Anomalies or gross errors;
- Non-anomalies or fine errors.

Given a distance function *d(.,.)* between two set of coordinates, and a threshold $\mathbf{E_r}$ in the related error, which represents a circle (or a sphere, in the 3-dimensional version) around the true source position, a localization error is classified as anomalous or gross error if $d(\mathbf{P_l},\mathbf{P_r}) > \mathbf{E_r}$**;** otherwise, it is classified as a non-anomalous or fine error. Thresholds for the discrimination between fine and gross errors will be different from *Accurate* localization tasks to *Rough* localization tasks: for instance, in a lecture scenario the threshold can be 50 cm for the *Accurate* localization task and 1 meter for the *Rough* localization task; on the other hand, in a meeting scenario, as just *Accurate* localization task makes sense, a reasonable unique threshold could be again 50 cm.

For what concerns the classification between gross and fine errors, one can also compute the localization rate P_{cor} ,as suggested in [8], which is defined as the number of fine errors (N_{FE}) over the total number of frames for which the localization system has produced a localization result (N_T):

$P_{cor} = N_{FE} / N_T$.

In the meeting scenario, this measurement will also be made speaker by speaker, in order to distinguish system performance for speakers not localized in favourite positions with respect to the far microphones (e.g. the speaker facing opposite to the NIST array will never produce a direct sound to any microphone of the array itself).

As previously highlighted, since error along *z*-coordinate seems to be less critical and more difficult to derive in an accurate way, the localization system performance will be evaluated by considering both 3 dimensions (*x*,*y*,*z*) and 2 dimensions (*x*,*y*). In both cases, every localization sequence will be represented by a list of (*x*,*y*) or (*x*,*y*,*z*) coordinate vectors, each of them corresponding to a given temporal interval, as discussed in the next section.

4.3 Temporal axis for evaluation

In this project, microphone signals are sampled at 44.1 kHz sampling frequency.

In principle, a speaker localization system may produce a set of coordinates at a very high rate, as the typical rate (100 Hz, i.e. every 10 ms) of an Automatic Speech Recognition (ASR) front-end or more, but in the given scenarios we feel that the adoption of a reduced rate in the range of, for instance, 1-10 Hz (which means that a set of coordinates will be produced every 100-1000 ms) is adequate. This choice should be consistent with a potential integration between audio processing and image processing systems for person localization and tracking purposes. In the following of this document, a temporal segment of 100 ms is assumed as a preliminary hypothesis[3].

Given 100 Hz rate, if a speaker localization system provides coordinates with a faster rate, the evaluation tool will average the coordinates, every 100 ms, on a 100 ms window centered around the given time instant. If the speaker localization system produces data with a slower rate, or is not able to produce a set of coordinates for some frames labelled as "one speaker" by the human labelers, the evaluation tool will classify those missing data as deletion errors. *Figure 2* provides examples of averaging, localization at exactly the frame rate, deletion and false alarm.

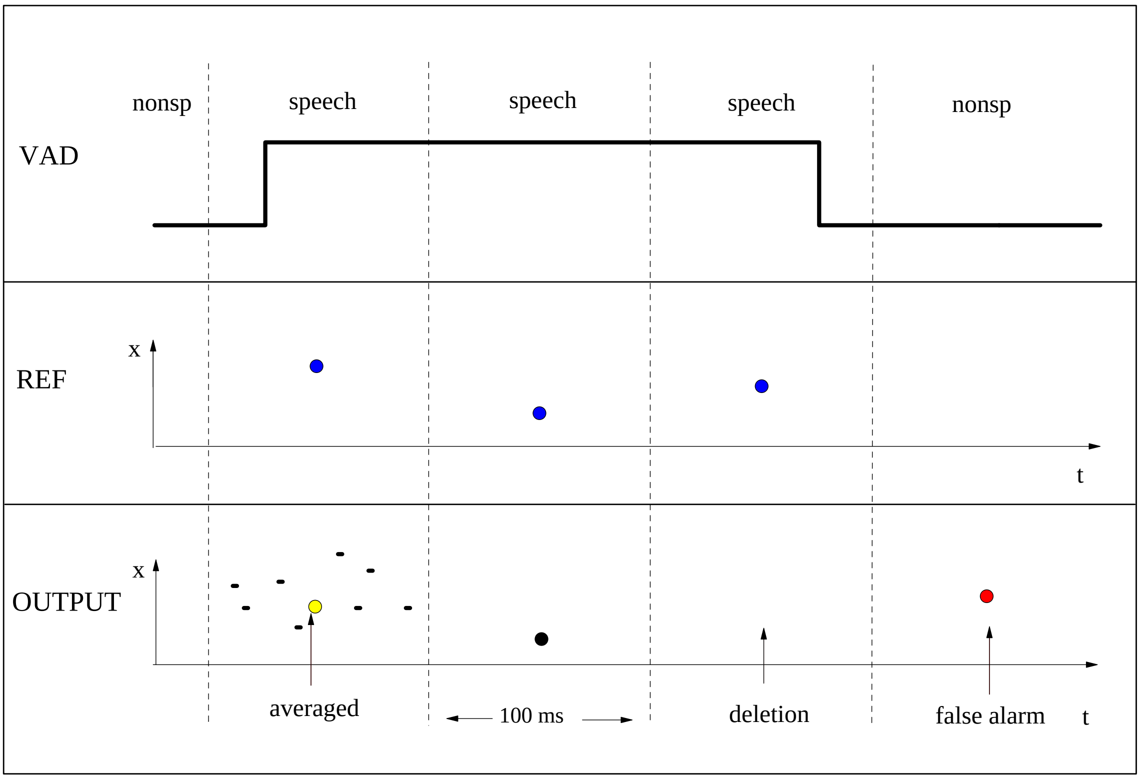

Figure 2: Examples of outputs of the localization system for the x coordinate: VAD is the bilevel information of the Speech Activity Detector, REF is the reference transcription of the x coordinate, OUTPUT shows the results of the localization system in the case of output at higher frame rate than 10 Hz, in the case of output at 10 Hz and in cases of deletion and false alarm, respectively.

Note that a speaker localization system can be conceived to produce anyway data every 100 ms (for instance deriving missing data by interpolation), in order to reduce or avoid deletion errors; in this way, the average localization error would probably increase. A speaker localization system will eventually be evaluated in terms of both good average localization error and low deletion error rate. The definition of a single evaluation feature derived from the latter two measures is still subject of discussion and depends on the real target application.

[3] As shown in Appendix B, in CHIL the first test experiments refered to a labeling rate corresponding to about 667 ms (i.e. 10/15 s), which derives both from the image sampling rate (equal to 1/15 s) provided by the given video recording systems and from the fact that manual labeling of the speaker position was done every 10 video frames.

4.4 Speech Activity Detection constraints

Any speaker localization system is based either on an implicit or on an explicit detection of a given event. So, first of all one could pose the following question: should the capability of localizing a speaker (or an acoustic source) depend or not on a preliminary common (to all the localization technologies) acoustic event detection processing. For instance, as in CHIL an activity is envisaged on the evaluation of speech activity detectors, one might use the best detector (or even the true detection reference coming from manual labeling) in order to select the input segments to the localization system. In this way, only the potentials of localization systems would be compared each other.

According to a preliminary discussion among the partners of CHIL working on speaker localization and tracking it was decided that the use by each partner of its SAD system (see [9]) to measure the impact of the entire SAD+speaker localization chain was interesting as well[4].

Although it may be considered as a minor comment, it is also worth noting that introducing a SAD pre-processor we also expect to have some false alarms due to it. However, we also expect that some automatic localization systems (but not all of them) will be conceived to derive an estimate of the speaker coordinates only when the SAD preprocessing module has detected a certain speech activity. Other localization systems may work in a different way, for instance producing a set of coordinates according to a given confidence measure of the localization reliability (and to a related thresholding). The latter situation would lead to other possible false alarms of the speaker localization system, not corresponding to frames detected as speech by SAD. Hence, the speaker localization systems are here evaluated in terms of a single level of <u>false alarm rate</u> (the extra alarms due to the localization system can eventually be derived on the basis of false alarm statistics computed on the preprocessing SAD system).

4.5 Competitive speakers

In order to fairly evaluate the speaker localization performance, at the moment the evaluation process takes into account only segments where one and only one talker is involved. On the basis of manual reference transcriptions, segments where <u>more talkers</u> (or one talker and at least one noise source) are active at the same time will not be counted in the final statistics.

4.6 Reference Transcriptions

In order to evaluate the given localization technologies, a very accurate labeling is needed. To summarize, reference transcriptions should consist in the following set of data for each frame (every 100-1000 ms), as follows:

- How many persons are talking in that specific instant: 0, 1, more than one.
- How many noise sources are active: 0, 1, more than one.
- Who is the speaker (or SpkId in meetings, and “Lecturer” vs “Audience” in the seminars)

[4] Note that in this way we would evaluate two capabilities of the system at once (to detect speech and to localize the speaker), i.e. an entire technology developed by the partner.. For this reason, the evaluation software was conceived in both ways.

- (x,y,z) coordinates of the speaker

The transcription files are produced partly extracting the related information both from what was transcribed for far-field ASR and for SAD evaluation purposes and from labelling of video recordings. To this purpose (as discussed in *Appendix A*) a specific software tool was developed to derive the reference transcription files from the XML based files produced for ASR evaluation purposes.

A possible extra transcription task can be necessary to obtain a consistent reference for speaker localization evaluation purposes. For instance, due to sound propagation effects, a time offset always holds between a close-talk and the far-field microphone recordings. For this reason, a preliminary check at this level is necessary. As there is no software to perform this check in an automatic way, this problem is just mentioned to recall that misleading results may come out due to this reason.

4.7 Evaluation tool

The evaluation software was developed to operate as follows.

The output of the speaker localization system (input of the evaluation tool) is provided as plain ascii text (one file for each evaluation segment, clearly named according to the data it refers to). Corresponding to each file, a reference file (see *Appendix A*) is available to derive the evaluation results.

The next two subsections, 4.7.1 and 4.7.2, report on an example of meeting scenario and an example of lecture scenario, respectively.

Each row of the output localization file must contain a time index (in seconds) and the estimated x, y, z coordinates (in mm). The evaluation system compares the localization results with the reference localization data and then provides a set of indexes describing the accuracy of the localization system (percentage of anomalous estimates, bias and standard deviation of non-anomalous estimates, deletion and false-alarm localization rates).

In the following examples an anomalous error threshold (E_r) of 50 cm is assumed in *Accurate* localization and of 1 meter in *Rough* localization. The time scale is obviously not realistic, with so many change of speakers in just one second, but it serves as an exemplification of the various situations. Here FAR denotes the False Alarm Rate (Number of false alarms/Number of frames with 0 speakers) and DR the Deletion Rate (Number of deleted frames/Number of frames with 1 speaker).

Note that the given example resembles the actual use of the current version of the evaluation software. For a meeting scenario the software is under development and the evaluation summary format will be updated accordingly.

4.7.1 Example of reference/input/output for a meeting scenario

Content of the reference file:

Frame time[s]	Number of active speakers	Number of active noise sources	Speaker ID	X coord. [mm]	Y coord. [mm]	Z coord. [mm]
0.0	1	0	Spk1	2200	3000	1200
0.1	0	1	-	-	-	-
0.2	1	0	Spk1	2150	2900	1200
0.3	1	0	Spk1	2150	2900	1200
0.4	2	1	-	ND	ND	ND
0.5	1	0	Spk2	3580	3500	1100
0.6	1	0	Spk2	3550	3450	1100
0.7	1	0	Spk2	3600	3500	1100
0.8	0	1	-	-	-	-
0.9	1	0	Spk1	2200	2950	1200
1.0	2	0	-	ND	ND	ND

Input (output of the speaker localization system):

Frame time [s]	X coord. [mm]	Y coord. [mm]	Z coord. [mm]
0.0	2000	3000	1200
0.2	2100	2950	1250
0.3	2100	2900	1250
0.4	2000	2950	1200
0.5	3500	3500	1050
0.6	3900	1500	1150
0.7	3500	3550	1150
0.8	1550	2100	1600

Evaluation output:

Frame time [s]	**Error [mm]**	**Classification**
0.0	200	Fine error
0.1	ND	No speaker
0.2	87	Fine error
0.3	71	Fine error
0.4	ND	Ignored (Multiple speakers)
0.5	94	Fine error
0.6	1982	Gross error
0.7	122	Fine error
0.8	ND	False Alarm
0.9	ND	Deletion
1.0	ND	Ignored (Multiple speakers)

Evaluation Summary:

	Spk1	**Spk2**	**Average**
$\mathbf{P_{cor}}$	3/3=1	2/3=0.66	5/6=0.83
Bias (x,y,z)	(-100,17,33)	(-90,25,0)	(-96,20,20)
RMSE [mm]	132	109	124
DR	1/4=0.25	0/3=0	1/7=0.14
FAR	-	-	1/2=0.5

4.7.2 Example of reference/input output for a lecture scenario

Content of the reference file:

Frame time[s]	Number of active speakers	Number of active noise sources	Speaker ID	X coord. [mm]	Y coord. [mm]	Z coord. [mm]
0	1	0	lecturer	4000	2000	1800
0.1	0	1	-	-	-	-
0.2	1	0	lecturer	3950	2500	1800
0.3	2	1	-	ND	ND	ND
0.4	1	0	audience	1200	1500	1200
0.5	1	0	lecturer	4100	2250	1700
0.6	0	2	-	-	-	-
0.7	1	0	audience	1200	1500	1100
0.8	2	0	-	ND	ND	ND
0.9	1	0	audience	1200	1500	1150
1.0	1	0	lecturer	3980	2380	1800
1.1	0	1	-	-	-	-
1.2	1	0	audience	1200	1500	1000
1.3	2	1	-	ND	ND	ND

Input (output of the speaker localization system):

Frame Time [s]	X (mm)	Y (mm)	Z (mm)
0	3960	1980	1800
0.2	4050	2390	1850
0.4	1100	1810	1100
0.5	4300	2800	1770
0.6	3000	3100	1150
0.7	1250	1700	1180
0.9	2000	2200	940
1.0	4100	2240	1860

Evaluation output:

Frame Time [s]	Error [mm]	Classification
0	45	Fine Error Lecturer
0.1	ND	No speaker
0.2	157	Fine Error Lecturer
0.3	ND	Ignored (Multiple speakers)
0.4	341	Fine Error Audience
0.5	589	Gross Error Lecturer
0.6	ND	False Alarm
0.7	221	Fine Error Audience
0.8	ND	Ignored (Multiple speakers)
0.9	1084	Gross Error Audience
1.0	194	Fine Error Lecturer
1.1	ND	No speaker
1.2	ND	Deletion Audience
1.3	ND	Ignored (Multiple speakers)

Evaluation Summary (given a timestep=0.1):

	Lecturer	**Audience**	**Overall**
P_{cor}	0.75 (=3/4)	0.67 (=2/3)	0.71 (=5/7)
Bias fine (x,y,z) [mm]	(60,-90,37)	(-25,255,-10)	(26,48,18)
Bias fine+gross (x,y,z) [mm]	(95,70,45)	(250,403,-77)	(161,213,-7)
RMSE fine [mm]	146	287	214
RMSE fine+gross [mm]	321	668	500
Deletion Rate	0.00 (=0/4)	0.25 (=1/4)	0.12 (=1/9)
False Alarm Rate	0.33 (=1/3)		
N. of loc. frames for error statistics	4	3	7
Total n. of output loc. frames = 8	**Reference duration = 1.3**	**Average frames/s = 6.15**	
Total n. reference frames = 14			

5. Calibration

5.1 Introduction

The accuracy of a speaker localization system is highly dependent on the precision with which the exact position of each sensor is known. In addition to a meticulous manual measurement of the coordinates of each microphone, we propose to use also a semi-automatic procedure of calibration useful to validate the geometric model of microphone arrangement [10, 11]. This operation can be accomplished by employing one or more loudspeakers in known positions and a test signal with appropriate characteristics. According to some preliminary experiments, the use of cheap PC loudspeakers would not alter the effectiveness of the proposed procedure.

The test waveform should facilitate the estimation of the relative delays between the signals acquired by the various microphones. A consistency check can be achieved in this way, and a calibration of the source localization system can be carried out on the test signals. Moreover, the possibility of estimating the room impulse response between a given source position and each microphone in the room (although in this case a hi-fi loudspeaker is surely required) is of great interest. In the fact, the knowledge of the impulse responses is advantageous to characterize the multipath propagation inside the room and to create realistic models for far-microphone signals acquired from real talkers. A more accurate modeling could be achieved by exploiting a talking head in place of the loudspeaker: however, this solution seems to be at the moment too complex to adopt in CHIL.

Finally, note that a specific issue that should not be neglected in a calibration phase concerns the dependence of the speed of sound on temperature. If not accounted for, this could introduce a bias that would directly affect the results of the localization procedure.

5.2 Chirp-like test signal

The time-stretched pulse proposed by Aoshima [12] and generalized in [13] is a chirp-like signal having a flat overall power spectrum, that enables a very accurate measurement of the acoustic impulse response. As a consequence of its extended time duration, this excitation can deliver a large amount of energy, while avoiding problems of dynamic range. The pulse is defined on the discrete frequency domain as the N-point sequence:

$$P(k) = \begin{cases} \exp(j2m\pi k^2 / N^2) & 0 \le k \le N/2 \\ P^*(N-k) & N/2 < k < N \end{cases}$$

where * denotes complex conjugate. The parameter *m* is an integer that determines the stretch of the pulse. The inverse DFT of *P(k)* is a chirp-like sequence *p(n)* that can be transduced by a loudspeaker into an acoustic signal.

A noteworthy characteristic of *p(n)* is that its autocorrelation is an almost perfect Dirac delta function. As a consequence, the sequence *y(n)*, acquired by a microphone when the loudspeaker generates this excitation, can be deconvolved easily by simply cross-correlating it with the original sequence *p(n)*. The result is the acoustic impulse response from the loudspeaker to the microphone. Apart from the contribution of frequency response of the loudspeaker, this is exactly the impulse response *h(t)* of the acoustic channel in the acquisition of a talker speaking at the same position of the loudspeaker.

5.3 Acquisition of calibration signals

Considering its properties, we suggest to use chirp-like signals to record calibration signals for all the microphones before each acquisition session of the real acoustic data foreseen in the CHIL scenarios.

We propose to use a calibration sequence consisting of two chirp-like signals (one from low-to-high frequencies and the other from high-to-low frequencies) and a short pulse. This sequence, reproduced by means of a loudspeaker, allows to estimate the impulse responses and to accurately calculate the propagation times from the loudspeaker to each microphone.

It is important to note that the quality of the used loudspeaker (its frequency response) may affect the accuracy of the calibration data. It is therefore recommended that at least a medium-quality device be employed.

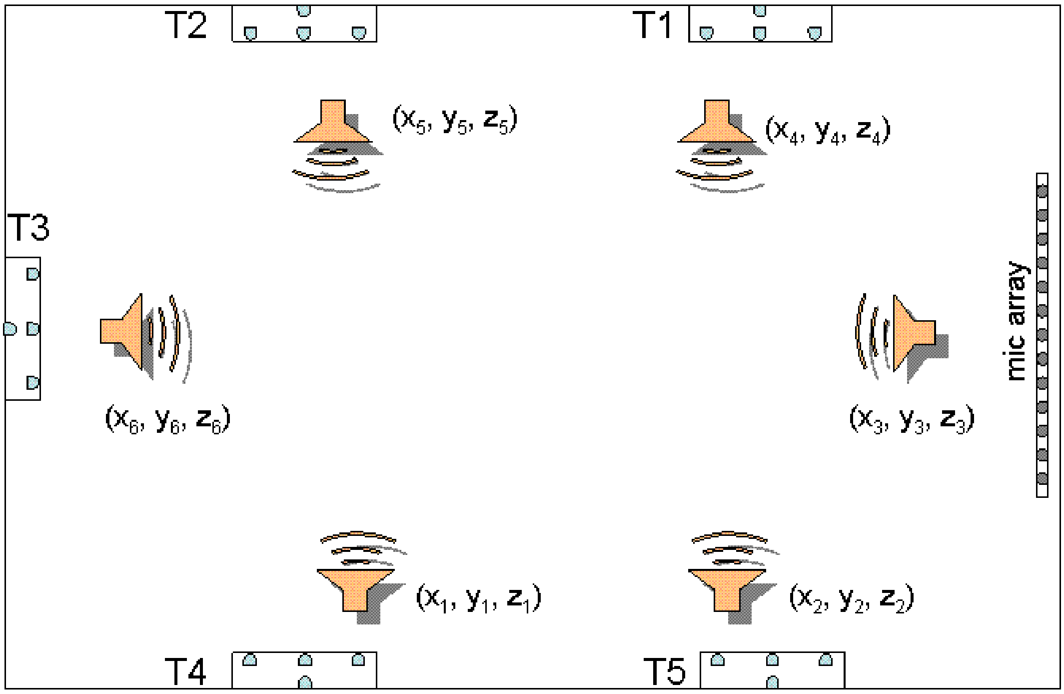

Figure 3: Example of a set of loudspeaker positions for the collection of calibration data inside a typical CHIL room.

The calibration sequence should be reproduced by the loudspeaker in a set of positions inside the room, at accurately measured coordinates. For each position (at least 5 or 6 in a typical CHIL room, uniformly distributed: see Figure 3) it is important to document the exact coordinates (x,y,z of the center of the emitting cone) and the orientation of the loudspeaker. All the synchronous channels will record the calibration signals while the sequence is being reproduced by the loudspeaker.

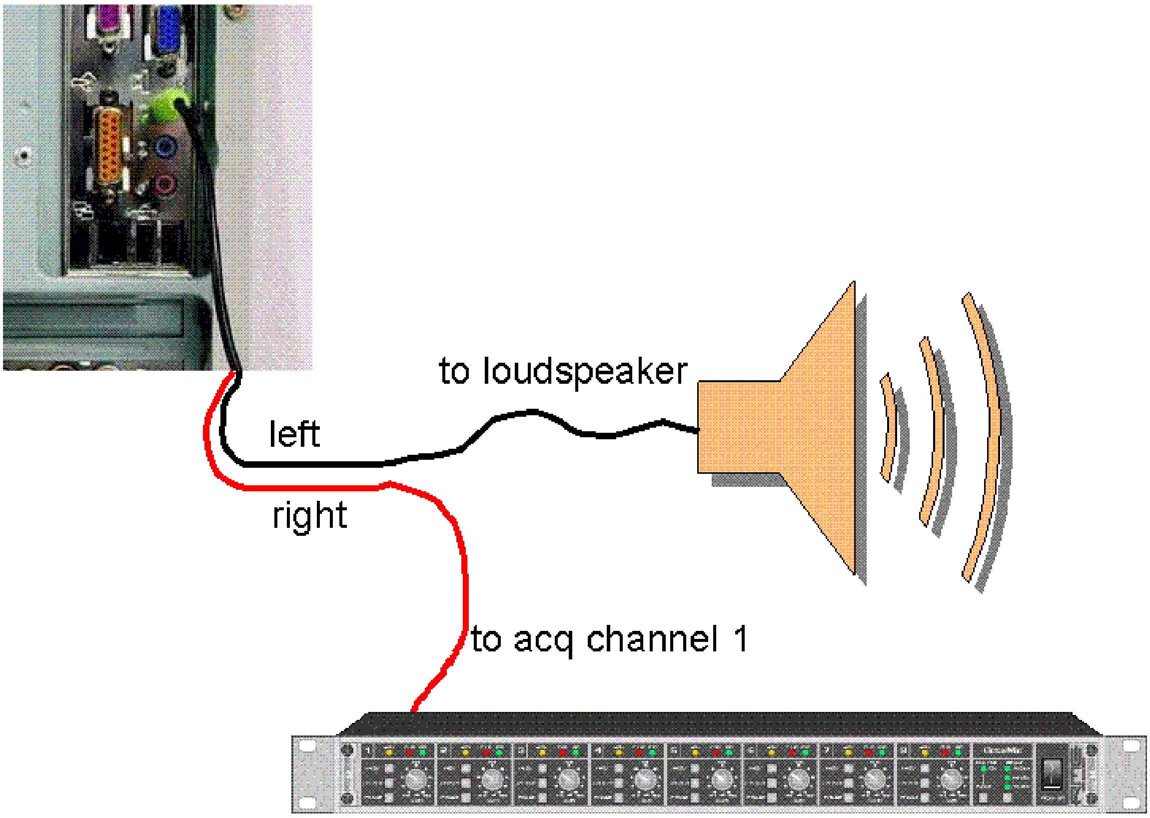

Figure 4: Example of connections to acquire a synchronous reference signal. The calibration sequence is being played through the loudspeaker at the same time redirected to one of the input channels.

In order to record a synchronous time reference of the emitted signal, the signal sent to the loudspeaker should be split and looped back into one of the acquisition channels. For example if the sequence is played by an audio-board, the stereo output of the board can be split by means of a suitable connection cable: the left channel will be sent to the loudspeaker while the right channel will be connected to one of the acquisition channels (see Figure 4).

The calibration sequence is available together with this document.

6. References

[1] J.R. Casas, R. Stiefelhagen, "***Multi-camera/multi-microphone system design for continuous room monitoring***", CHIL Deliverable 4.1, July 2004.

[2] M. Brandstein, D. Ward, eds., ***"Microphone Arrays: Signal Processing Techniques and Applications"***, Springer-Verlag, Berlin, 2001.

[3] C.H. Knapp, C. Carter, ***"The generalized correlation method for estimation of time delay"***, IEEE Trans. ASSP, 24, pp 320-327, 1976.

[4] M. Omologo, P. Svaizer, ***"Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique"***, Proc. IEEE ICASSP 1994.

[5] M. Omologo, P. Svaizer *__"Acoustic Source Localization in Noisy and Reverberant Environment using CSP Analysis"__*. Proc. IEEE ICASSP 1996.

[6] J. Chen, J. Benesty, Y. Huang, *__"Robust Time Delay Estimation Exploiting Redundancy Among Multiple Microphones",__* IEEE Trans. on SAP, vol. 11, no. 6, 2003.

[7] B. Champagne, S. Bédard, A. Stéphenne, *__"Performance of time delay estimation in the presence of room reverberation"__*, IEEE Trans. SAP, vol. 4, pp. 148-152, 1996.

[8] T. Nishiura, T. Yamada, S. Nakamura, K. Shikano , *__"Localization of multiple sound source based on a CSP analysis with a microphone array__* ", Proc IEEE ICASSP 2000.

[9] D. Macho, J. Padrell, C. Nadeu, A. Temko, "*__Metrics for Evaluation of Speech Detection Technology__*", CHIL document, May 2004.

[10] J.M. Sachar, H.F. Silverman, W.R. Patterson III, *__"Position calibration of large-aperture microphone arrays"__*, Proc. IEEE ICASSP 2002.

[11] V.C. Raykar, R. Duraiswami, *__"Automatic position calibration of multiple microphones"__*, Proc. IEEE ICASSP 2004.

[12] M. Aoshima, *__"Computer-generated pulse signal applied for sound measurement"__*, JASA, vol. 69, 1981.

[13] Y. Suzuki, F. Asano, H.Y. Kim, T. Sone, "*__An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses__*", JASA, vol. 97, 1995.

7. Appendix A: Software for conversion from XML labeling to reference files

This software is necessary when one has to derive the localization reference file to use as input to the evaluation software discussed in Section 4. In fact, that evaluation software was not conceived to process a XML reference labeling file as input.

The name of the software is “make_reference_file.pl” and it consists in a PERL program which creates a relationship between time-position data in the 3D coordinates produced by manual labeling file and time-speech data in the transcription file.

Software input
The software loads the localization data, a tab-delimited table with a time (in seconds) column and three position column (x, y and z, in millimeters) in four arrays. For computational efficiency the time column must be in chronological order: every time found in this file is a time frame which starts at the indicated time and ends at the next time. Then the software loads the first part of the transcription file, storing speakers' data in a hash table.
Finally, it parses the rest of the transcription file, looking for noise sources and speech activity.
Whenever one of the following situation is encountered, the parser stores the data in a noise's or speech's array:
- At a turn begin or synchronization tag, the current time is updated;
- At a turn begin tag, the parser marks every frame from the speechstart to the speechend times with the current speaker indication;
- At a silence begin tag, the parser removes every speaker indication in the frames from the current time until a silence end tag;
- At a noise begin tag, the parser marks every frame with a noise source from the current time until a noise end tag;
- At an instantaneous silence tag, the parser removes every speaker indication from the current frame;
- At an instantaneous noise tag, the parser marks the current frame with a noise source.
Since in the transcription file time markers may not coincide exactly with the beginning of time frames from the localization file, the time frame considered is always the one in which the transcription time falls.

Problems found with transcription data
- In the transcription file the silence marker is used as instantaneous, even when it is obvious that it refers to an entire period of time between two synchronizaton tags. This results in an instantaneous silence instead of a long silence.
- Many times in the transcription file the noise begin tags and noise end tags are used instead of instantaneous noise tags, or the noise end tag is omitted. This results is a never ending noises;
- The starting time in the two given files may differ.

Partial solutions
- Instantaneous silence tag is considered only a silence begin tag;
- Turn end tag and synchronization tag is considered also silence end tag;

- Turn end tag is considered also an all noises end tag;
- From the localization file's times the program subtracts the first localization time (thus beginning with time 0 as in the transcription file) and sums an offset parameter specified by the user.

Unsolved problems
- There are too few synchronization tags in the transcription file. Many noises are simply put inside the speech segment between two very far synchronization tags, and locating them at their right time is not possible. Therefore, in the output file the first frame of every speech segment is very noisy;
- The solution for missing noise end tags works partially and may produce too long noises;
- Speakers very often do not speak together; there are few frames with more than one speaker and they are always located between two turns.

Program output
The program returns a tab-delimited table with the frame's starting time, the number of speakers, the number of noise sources, the speaker ID (if the speaker is one), the X Y Z coordinates (if the speaker is one; ND if the speakers are more than one; - if there is no speaker).

Program usage
make_reference_file.pl [-*offset=OFFSET_SECONDS] localization_file transcriptions_file.trs*

8. Appendix B: Example of input-output files for a real seminar data set

An example of use of the given software is presented in the following. The data were extracted from the July 21st seminar.

The software presented in Appendix A was initially applied to the files **File.trs** (Reference XML Transcriptions) and **File.3dl** (time-labeled video coordinates). A segment of both files is reported in the following.

Afterwards, the evaluation software presented in Section 4.7 was applied to the resulting **File.ref** and **irst.loc**, that is the output of a speaker localization system (developed at ITC-irst). Also for these files, a short segment is reported to give a rough idea of the file formats.

A summary of the results, obtained on the entire seminar, is reported at the end of the appendix.

File.trs

```
<Sync time="0"/>
<Event desc="sil" type="noise" extent="instantaneous"/>
<Sync time="13.982"/>
<Event desc="de" type="language" extent="begin"/>
<Event desc="door" type="noise" extent="instantaneous"/>
<Sync time="19.269"/>
<Event desc="door" type="noise" extent="instantaneous"/>
<Event desc="fst" type="noise" extent="instantaneous"/>
<Event desc="pap" type="noise" extent="instantaneous"/>
<Event desc="de" type="language" extent="end"/>
 yo okay
<Event desc="b" type="noise" extent="instantaneous"/>
 % today for the final we are talking about
<Event desc="b" type="noise" extent="instantaneous"/>
 body tracking for
<Event desc="^^" type="lexical" extent="instantaneous"/>
 gesture interfaces that is how can you use gestures to control a computer
<Event desc="b" type="noise" extent="instantaneous"/>
<Event desc="mn" type="noise" extent="previous"/>
not to move a cursor or
<Event desc="b" type="noise" extent="instantaneous"/>
 do other useful things$
<Sync time="79.114"/>
…
```

File.3dl

```
1090418497.751 1214.54 4180.58 1702.75
1090418498.417 1201.43 4157.12 1707.78
1090418499.084 1200.92 4157.04 1712.83
1090418499.750 1202.30 4160.50 1711.34
1090418500.417 1200.26 4146.99 1711.89
1090418501.083 1206.12 4158.05 1713.64
1090418501.750 1207.55 4148.00 1711.07
1090418502.416 1219.81 4153.51 1706.45
1090418503.083 1209.12 4162.80 1705.24
1090418503.749 1203.56 4151.63 1718.81
```

```
1090418504.416 1204.73 4155.96 1718.92
1090418505.082 1218.53 4137.53 1702.55
1090418505.749 1224.90 4146.92 1702.67
1090418506.415 1228.19 4143.24 1701.50
1090418507.082 1229.53 4161.49 1706.67
1090418507.748 1227.56 4172.96 1704.26
1090418508.415 1222.66 4156.81 1703.60
1090418509.081 1159.36 4271.09 1730.86
1090418509.748 1151.39 4347.23 1717.92
1090418510.414 1144.35 4335.82 1722.64
1090418511.081 1150.69 4323.87 1713.10
1090418511.747 1154.32 4396.64 1705.90
1090418512.414 1193.23 4408.38 1699.76
1090418513.080 1187.20 4334.03 1721.91
1090418513.747 1215.13 4348.79 1721.44
1090418514.413 1517.95 4334.24 1695.60
1090418515.080 1586.07 4298.22 1700.61
1090418515.746 1522.08 4308.61 1704.35
1090418516.413 1522.60 4297.05 1704.26
1090418517.079 1510.86 4283.77 1698.02
1090418517.746 1504.90 4306.93 1696.61
1090418518.412 1505.07 4292.70 1701.35
1090418519.079 1541.11 4309.25 1692.83
1090418519.745 1493.60 4349.49 1702.45
1090418520.412 1259.27 4376.75 1732.82
1090418521.078 1153.48 4345.26 1727.85
1090418521.745 1155.41 4356.11 1724.66
1090418522.411 1149.01 4345.61 1726.42
1090418523.078 1133.45 4347.38 1731.80
1090418523.745 1132.78 4336.95 1725.80
1090418524.411 1167.38 4352.65 1723.71
1090418525.078 1174.41 4356.23 1717.66
1090418525.744 1173.76 4361.85 1714.62
1090418526.411 1146.35 4379.11 1727.63
1090418527.077 1263.81 4370.07 1716.94
1090418527.744 1390.93 4307.59 1722.94
1090418528.410 1242.53 4369.40 1727.34
1090418529.077 1190.43 4326.58 1725.84
1090418529.743 1211.95 4359.49 1721.21
1090418530.410 1222.67 4369.13 1711.30
1090418531.076 1217.34 4363.20 1727.15
1090418531.743 1213.35 4381.78 1732.60
1090418532.409 1204.28 4363.53 1718.26
1090418533.076 1202.02 4356.61 1727.60
1090418533.742 1195.17 4363.04 1724.64
1090418534.409 1192.58 4345.97 1718.10
1090418535.075 1194.01 4339.78 1726.70
1090418535.742 1208.38 4320.60 1720.22
1090418536.408 1219.26 4353.44 1715.42
1090418537.075 1181.29 4344.51 1712.30
1090418537.741 1116.85 4239.82 1723.57
```

…

File.ref

```
 0.000      1     3     audience    1214.54     4180.58     1702.75
 0.666      1     1     audience    1201.43     4157.12     1707.78
 1.333      1     1     audience    1200.92     4157.04     1712.83
 1.999      1     0     audience    1202.30     4160.50     1711.34
 2.666      1     1     audience    1200.26     4146.99     1711.89
 3.332      1     0     audience    1206.12     4158.05     1713.64
```

3.999	1	1	audience	1207.55	4148.00	1711.07
4.664	1	1	audience	1219.81	4153.51	1706.45
5.332	1	0	audience	1209.12	4162.80	1705.24
5.998	1	0	audience	1203.56	4151.63	1718.81
6.664	1	0	audience	1204.73	4155.96	1718.92
7.331	1	1	audience	1218.53	4137.53	1702.55
7.998	1	0	audience	1224.90	4146.92	1702.67
8.664	1	0	audience	1228.19	4143.24	1701.50
9.331	1	1	audience	1229.53	4161.49	1706.67
9.996	1	1	audience	1227.56	4172.96	1704.26
10.664	1	0	audience	1222.66	4156.81	1703.60
11.330	1	1	audience	1159.36	4271.09	1730.86
11.996	2	1	-	ND	ND	ND
12.663	1	0	lecturer	1144.35	4335.82	1722.64
13.330	1	1	lecturer	1150.69	4323.87	1713.10
13.996	1	0	lecturer	1154.32	4396.64	1705.90
14.663	1	1	lecturer	1193.23	4408.38	1699.76
15.328	1	0	lecturer	1187.20	4334.03	1721.91
15.996	1	0	lecturer	1215.13	4348.79	1721.44
16.662	1	0	lecturer	1517.95	4334.24	1695.60
17.328	1	0	lecturer	1586.07	4298.22	1700.61
17.995	1	0	lecturer	1522.08	4308.61	1704.35
18.662	1	0	lecturer	1522.60	4297.05	1704.26
19.328	1	1	lecturer	1510.86	4283.77	1698.02
19.995	1	1	lecturer	1504.90	4306.93	1696.61
20.661	1	0	lecturer	1505.07	4292.70	1701.35
21.328	1	0	lecturer	1541.11	4309.25	1692.83
21.993	1	0	lecturer	1493.60	4349.49	1702.45
22.661	1	0	lecturer	1259.27	4376.75	1732.82
23.327	1	0	lecturer	1153.48	4345.26	1727.85
23.993	1	1	lecturer	1155.41	4356.11	1724.66
24.660	2	1	-	ND	ND	ND
25.327	1	1	audience	1133.45	4347.38	1731.80
25.993	1	1	audience	1132.78	4336.95	1725.80
26.660	1	1	audience	1167.38	4352.65	1723.71
27.327	1	1	audience	1174.41	4356.23	1717.66
27.993	1	1	audience	1173.76	4361.85	1714.62
28.660	1	2	audience	1146.35	4379.11	1727.63
29.325	1	1	audience	1263.81	4370.07	1716.94
29.993	1	1	audience	1390.93	4307.59	1722.94
30.659	1	1	audience	1242.53	4369.40	1727.34
31.325	1	1	audience	1190.43	4326.58	1725.84
31.992	1	1	audience	1211.95	4359.49	1721.21
32.659	1	1	audience	1222.67	4369.13	1711.30
33.325	1	1	audience	1217.34	4363.20	1727.15
33.992	1	1	audience	1213.35	4381.78	1732.60
34.657	1	1	audience	1204.28	4363.53	1718.26
35.325	1	1	audience	1202.02	4356.61	1727.60
35.991	1	1	audience	1195.17	4363.04	1724.64
36.657	1	1	audience	1192.58	4345.97	1718.10
37.324	1	1	audience	1194.01	4339.78	1726.70
37.991	1	1	audience	1208.38	4320.60	1720.22
38.657	1	1	audience	1219.26	4353.44	1715.42
39.324	1	1	audience	1181.29	4344.51	1712.30
39.990	1	1	audience	1116.85	4239.82	1723.57

…

Irst.loc

```
1.5790 4780.1103 4375.4387 1700
1.6718 4750.0452 4333.3959 1700
```

```
2.7864 4882.1094 4174.1125 1700
2.8793 4866.7159 4175.6593 1700
8.5449 4116.2935 4050.8229 1700
8.6378 4118.1202 4052.8479 1700
15.7896 1866.7382 5266.3619 1700
15.9753 1870.0731 5263.9021 1700
16.0682 1870.7747 5266.6482 1700
18.7617 2051.8570 5229.8227 1700
20.2478 1871.5584 5277.4429 1700
26.2850 1891.5015 4805.6446 1700
26.3779 1893.6504 4794.5432 1700
26.6565 1863.2854 4781.5296 1700
27.9568 1979.1335 4790.5754 1700
28.0497 1981.2270 4787.0496 1700
28.1426 1996.5527 4771.4250 1700
28.2355 2001.8051 4768.0612 1700
30.3717 1272.9707 5096.0096 1700
32.2293 1627.3651 5171.4025 1700
32.9723 1580.4068 4988.5562 1700
33.2510 1565.9440 4828.9623 1700
33.3439 1567.6669 4825.6778 1700
34.0869 1837.0540 4485.7508 1700
34.1798 1850.2881 4482.3812 1700
35.8516 2174.7538 4301.7197 1700
36.2231 2179.1969 4313.4681 1700
36.6875 2184.5227 4289.1108 1700
37.8950 2030.8487 4505.3843 1700
39.5668 2042.4003 4505.5560 1700
39.7526 2011.1335 4490.4007 1700
39.8454 2026.8443 4476.7478 1700
...
```

Results (regarding the entire seminar) derived by running the evaluation software(*) as follows:

"evaluation.exe -reference File.ref -inputFile Irst.loc -evalOutput Out.eval -evalSummary Out.sum -thresholdLecturer 500 -thresholdAudience 1000 -timestep 0.667 -maxN 0"

Out.sum:

```
                               Lecturer               Audience               Overall
Pcor                           0.79                   0.28                   0.63
Bias fine (x,y,z)[mm]          (59,-34,-18)           (-25,-19,33)           (46,-32,-11)
Bias fine+gross (x,y,z)[mm]    (95,-176,-15)          (300,-2287,-15)        (162,-858,-15)
RMSE fine [mm]                 233                    376                    259
RMSE fine+gross [mm]           777                    2822                   1727
Deletion rate                  0.57                   0.67                   0.61
False Alarm rate                                                             0.47
Loc.frames for error statistics 909                  43                     952
Total n. of output loc.frames=3546    Reference Duration=2178.82   Average Frames/sec=1.63
Total n. of reference frames =3265
```

(*) The most recent release of the evaluation software allows to define the parameter "maxN", that represents the number of noise events that will be neglected at each frame. This feature was included to increase the number of localization frames for error statistics (i.e., reducing the number of Ignored frames), since at that time the labelling files referred to seminars recorded in November 2004 were characterized by long segments labelled with at least one active noise source (although not active during most of that segment).